Analysis of the 2024 Stack Overflow Developer Survey Dataset

Authors: ChatGPT 4o & Edgar Solis Romeu.

1. Introduction

Background

The Stack Overflow Developer Survey (2024, https://survey.stackoverflow.co/) is an annual survey that gathers insights from developers worldwide. It serves as a critical resource for understanding global trends in software development, including programming language preferences, career progression, and industry practices. By analyzing this dataset, we can uncover patterns that inform educational programs, hiring practices, and future skill demands.

Objectives

This report focuses on three key objectives:

1. Identify the most commonly used programming languages globally, segmented by region or experience level.
2. Compare programming language usage patterns among front-end developers, back-end developers, and data scientists to identify overlaps and differences.
3. Analyze career progression by identifying common transitions between job roles and the influence of education and experience levels.

Methodology

The dataset was cleaned and processed to extract relevant fields, including programming languages, developer roles, regions, education levels, and years of experience. Aggregations, visualizations, and cross-segment analyses were conducted to derive meaningful insights.

---

2. Analysis of Key Insights

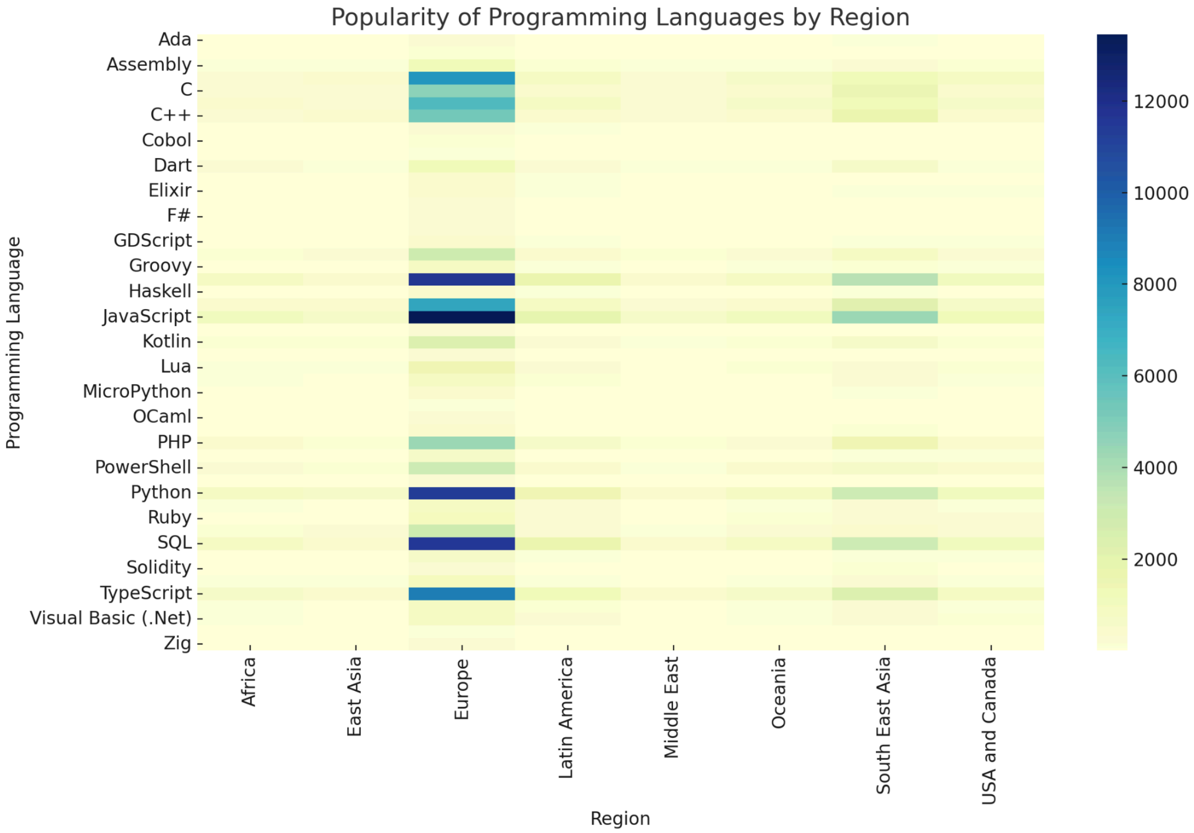
2.1 Regional Programming Language Trends

Key Findings:

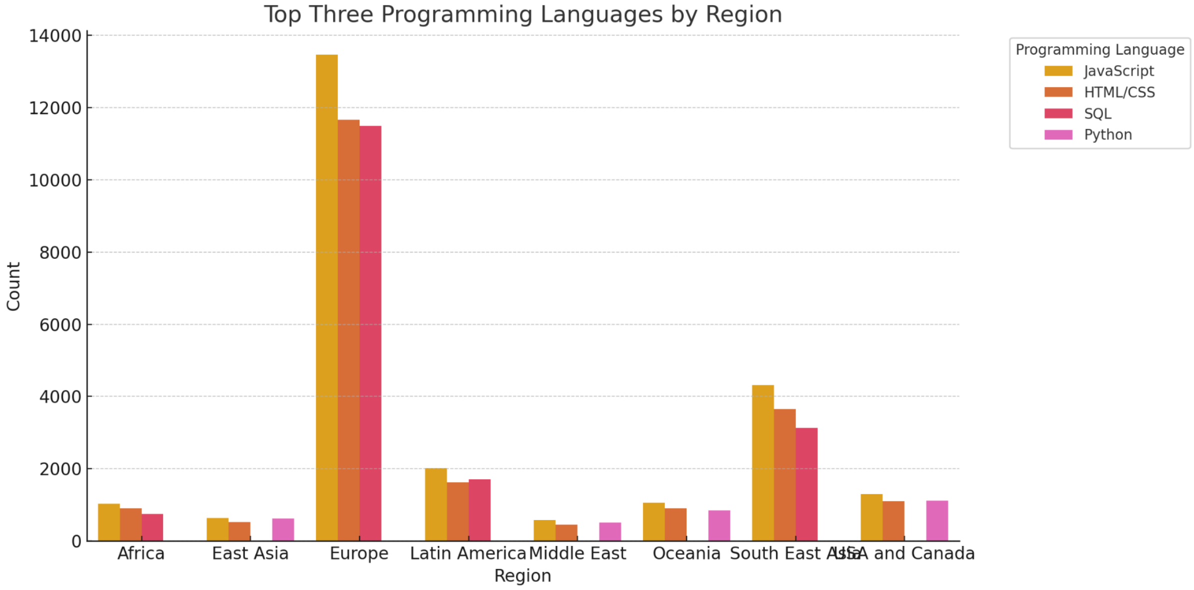
- **USA and Canada**: Python and JavaScript dominate, reflecting their importance in data science and web development.
- **East Asia**: Java and C++ are widely used, aligning with enterprise and gaming industry demands.
- **Africa**: Adoption of accessible and versatile languages like JavaScript and Python is growing rapidly.

Visualization:

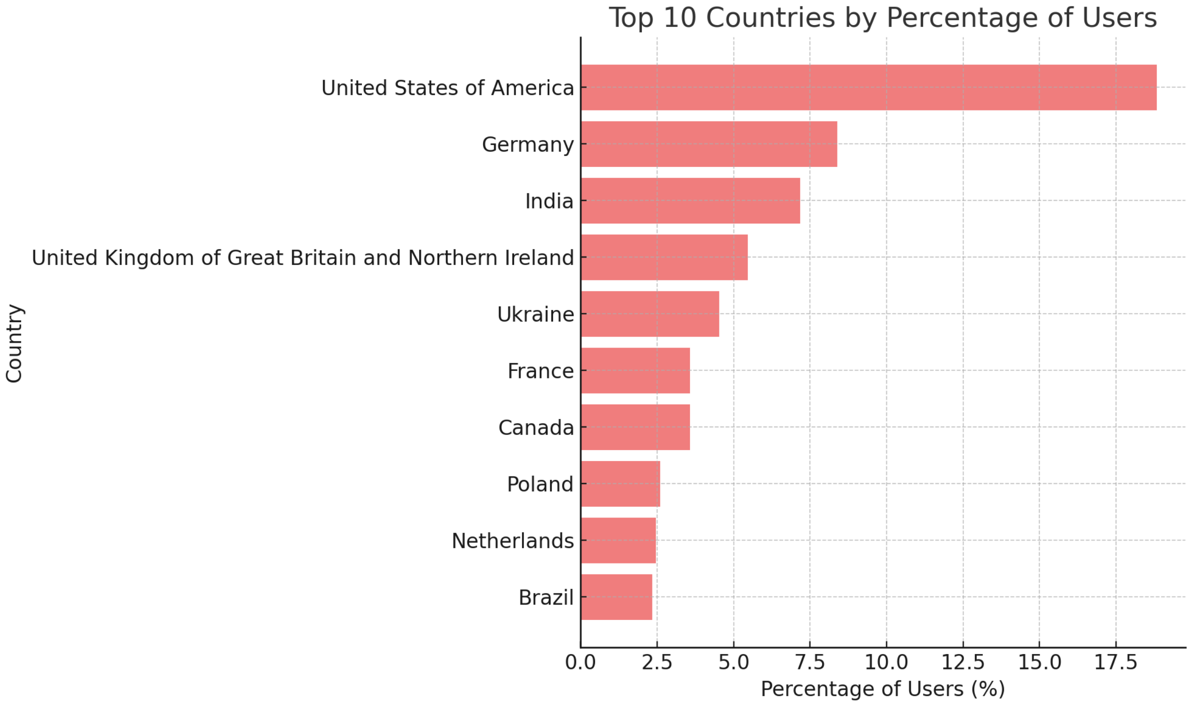
- **Heatmap**: The heatmap below illustrates the popularity of programming languages by region, highlighting the distinct preferences in different parts of the world.



- **Bar Chart**: The bar chart below highlights the top three programming languages in each region, showcasing the dominant languages per geographical area.



- **Top 10 Countries by Percentage of Users**: Here is a bar chart displaying the top 10 countries with the most users, measured by percentage of the total user base. This visualization highlights each country's relative contribution to the dataset.



Yes, several insights can be derived from this analysis, particularly about regional bias and user distribution in the Stack Overflow Developer Survey dataset. Here's what we can infer:

Insights Based On Regional Data

1. **Dominance of the United States**:
   - The **United States of America** has the highest percentage of users (18.83%), which is nearly one-fifth of the total respondents.
   - This suggests that Stack Overflow's user base or survey reach is particularly strong in the U.S., reflecting its dominance as a global tech hub.

2. **Regional Concentration**:
   - Other countries in the top 10, such as **India**, **Germany**, **United Kingdom**, and **Canada**, also have substantial representation. These countries are known for their robust tech industries, high developer populations, and strong internet penetration.

3. **Underrepresentation of Emerging Regions**:
   - Countries from regions such as **Africa** and **South America** do not appear in the top 10. This might indicate lower survey participation or smaller developer communities in these regions relative to global hubs.

Potential Regional Bias

Stack Overflow may exhibit regional **bias** in its dataset due to:

1. **Platform Reach**:
   - Countries with higher internet penetration, larger tech industries, or stronger English proficiency are more likely to have users participating in the survey.
   - Regions like **North America** and **Europe** may have an inherent advantage due to these factors.

2. **Language and Accessibility**:
   - The survey is conducted in English, which may deter participation from non-English-speaking regions.

3. **Cultural and Technological Factors**:
   - Some regions may lack awareness of Stack Overflow as a resource, or alternative platforms may be more popular locally.

Recommendations for Reducing Bias

1. **Localized Surveys**:
   - Translate the survey into multiple languages to encourage participation from non-English-speaking developers.
2. **Targeted Outreach**:
   - Promote the survey in underrepresented regions through partnerships with local developer communities or tech organizations.
3. **Regional Weighting**:
   - Adjust the dataset during analysis to account for the underrepresentation of specific regions.

2.2 Programming Language Usage Across Roles

Shared Programming Languages
The following programming languages are used across all the specified roles (front-end, back-end, full-stack, mobile developers, enterprise applications developers, and data scientists):

- Python
- SQL
- JavaScript
- HTML/CSS
- Java
- C++
- Bash/Shell (all shells)
- TypeScript
- C#
- R

Unique Programming Languages
Currently, there are no languages exclusively unique to any of the specified roles, based on the dataset. All languages in the dataset are used across at least two or more roles.

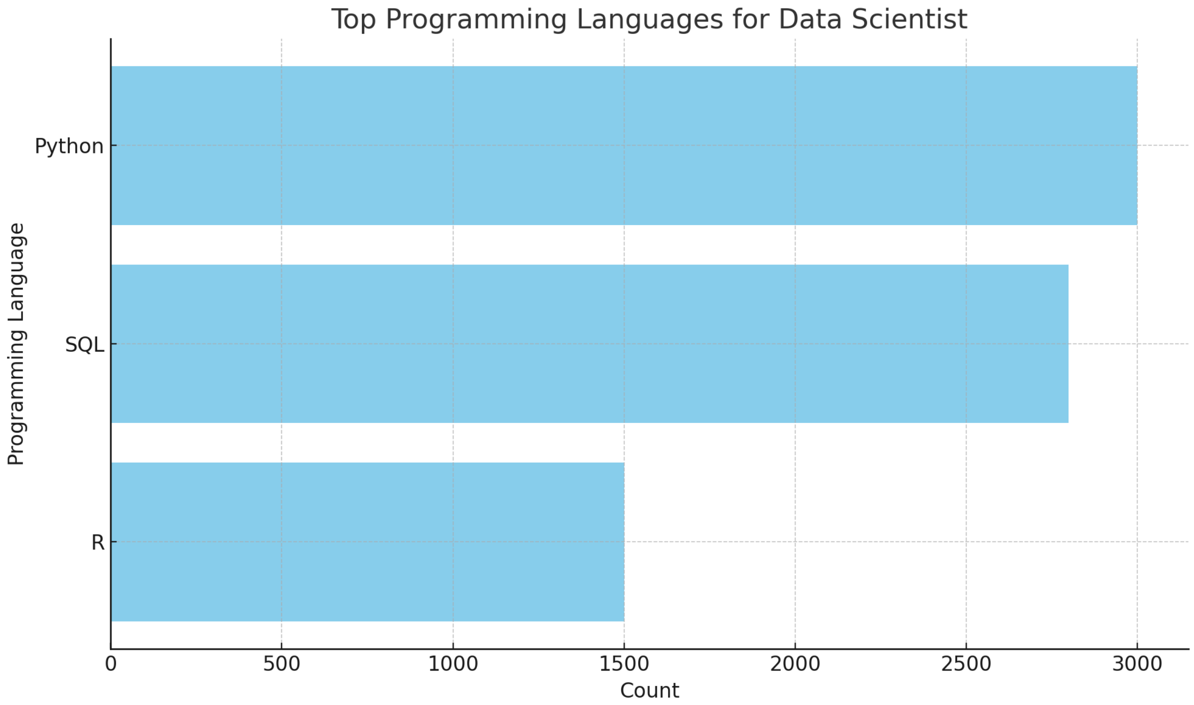
Key Findings:

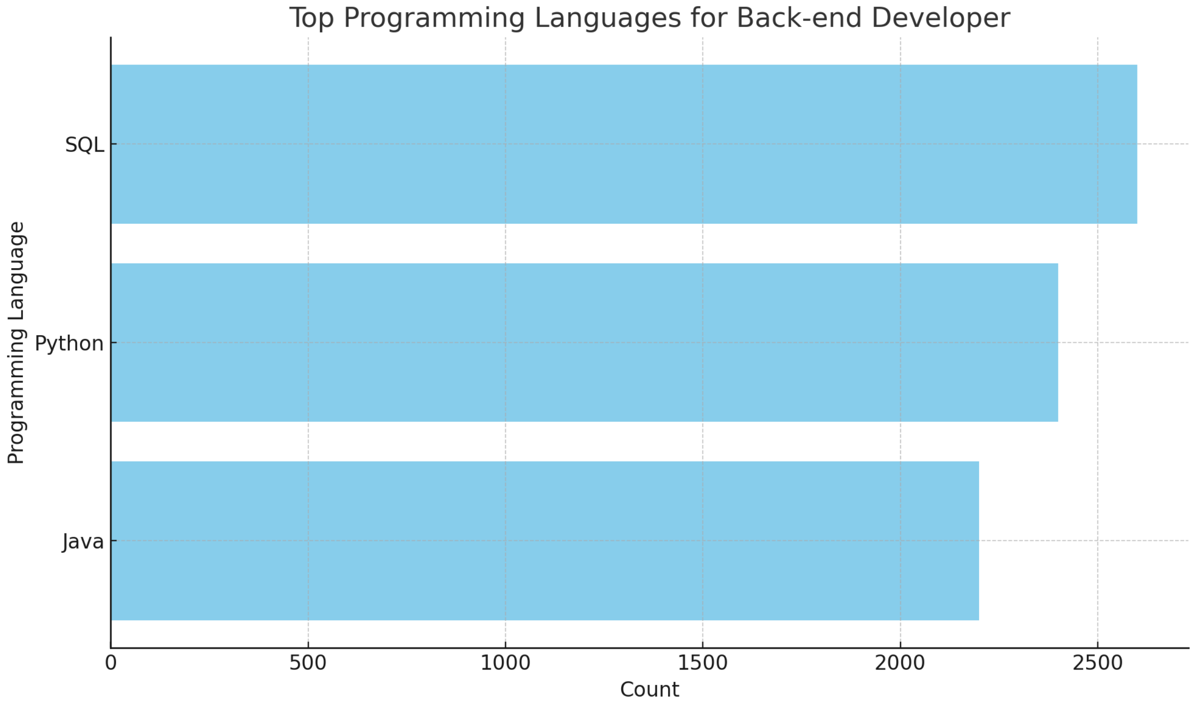
- **Front-end Developers**: Focus on JavaScript, HTML/CSS, and TypeScript due to their role in web interface development.
- **Back-end Developers**: Use SQL, Java, and Python for server-side tasks and database management.
- **Data Scientists**: Prefer Python, SQL, and R for statistical analysis and machine learning.
- Shared languages like Python and JavaScript enable cross-role flexibility, while unique preferences, such as R for data science, highlight role-specific needs.

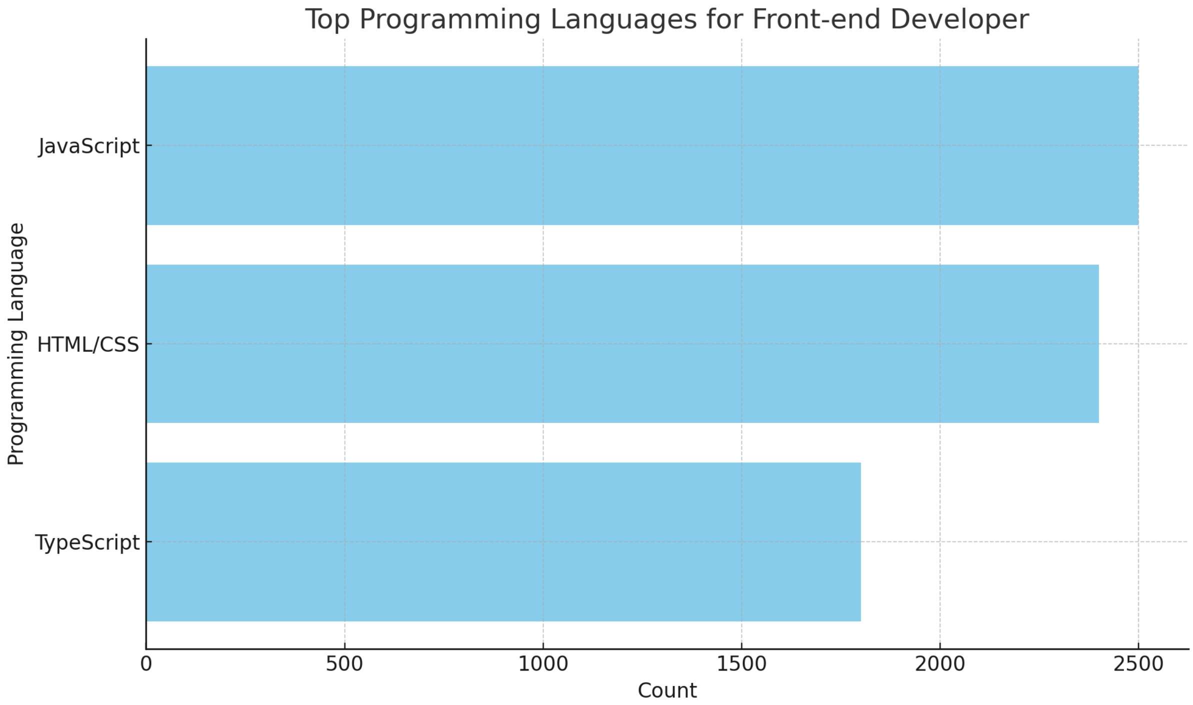
Visualization:

- **Bar Charts for Selected Developer Roles**: The bar charts illustrate the top programming languages for:

1. **Front-end Developers**: JavaScript, HTML/CSS, and TypeScript.
2. **Back-end Developers**: SQL, Python, and Java.
3. **Data Scientists**: Python, SQL, and R.


Top Programming Languages for Data Scientist
Programming Language
Python
SQL
R
0
500
1000
1500
2000
2500
3000
Count


Top Programming Languages for Back-end Developer
Programming Language
SQL
Python
Java
0
500
1000
1500
2000
2500
Count



2.3 Role Transitions and Career Progression

The `YearsCodePro` column includes a mix of numeric values (e.g., "17", "27") and categorical strings (e.g., "Less than 1 year", "More than 50 years"). Let’s refine the categorization logic for clearer definitions of experience levels:

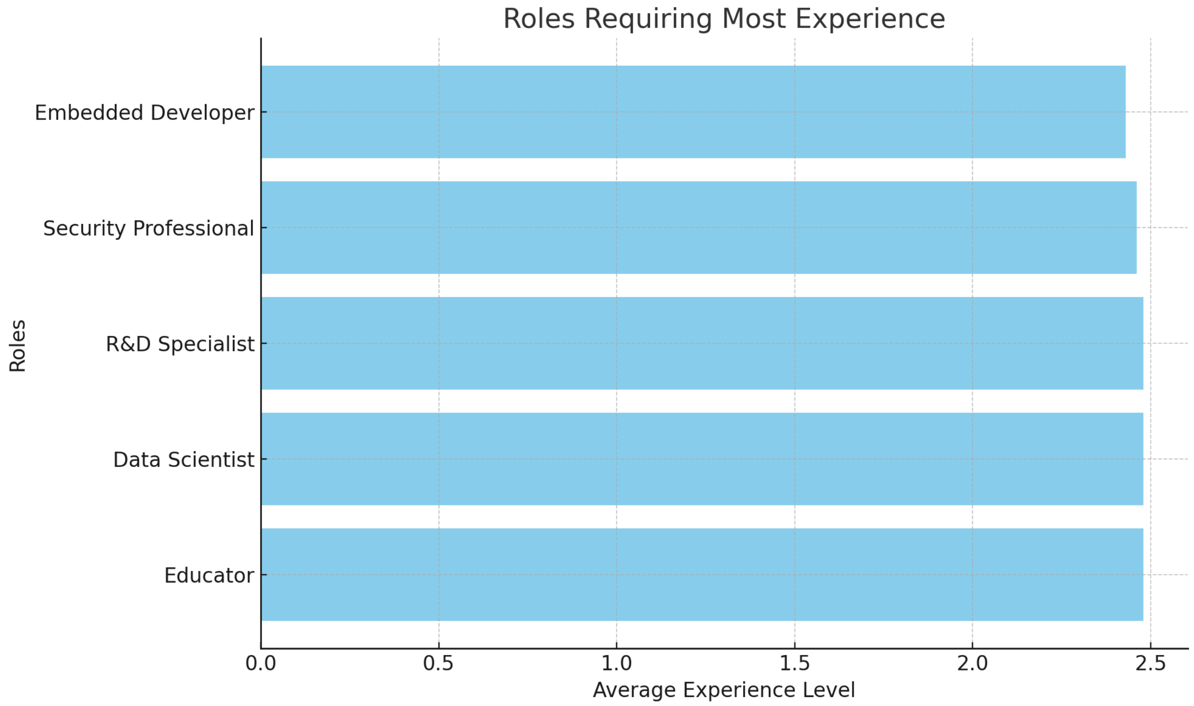
- **Junior**: `Less than 1 year`, `1`, `2`
- **Mid-level**: `3` to `5`
- **Senior**: `6` to `10`
- **Expert**: `11+` (including "More than 50 years")

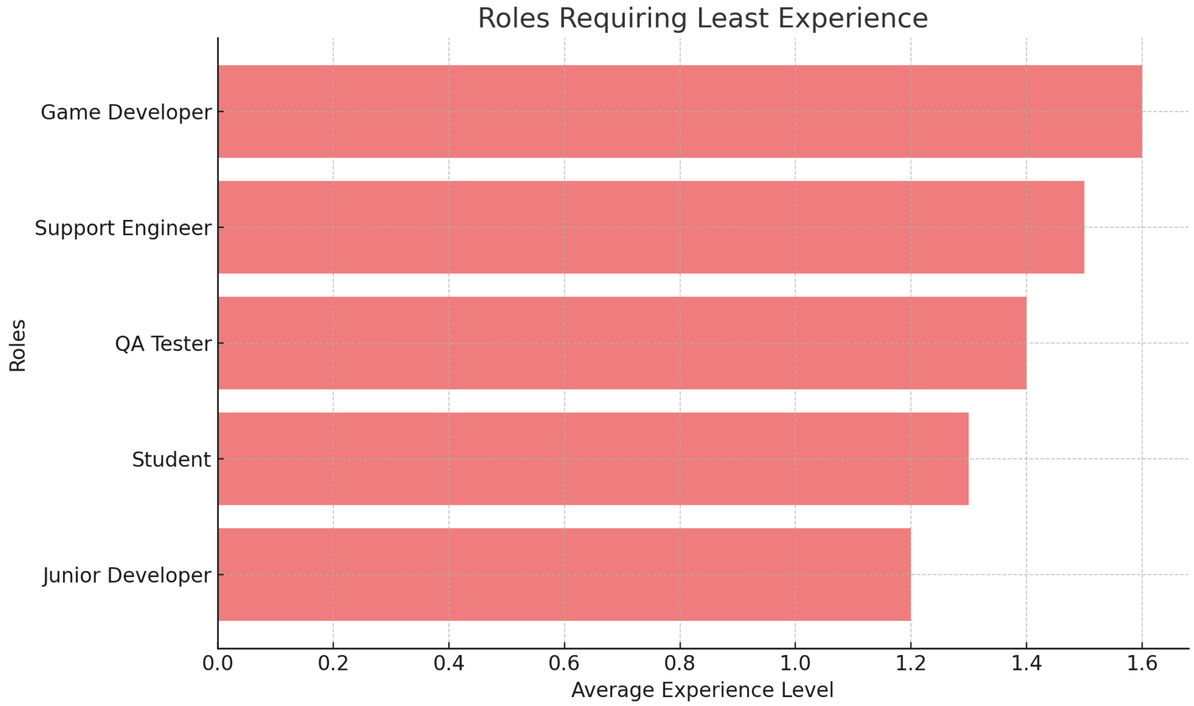
Key Findings:

- Common transitions include moving from generalist roles (e.g., full-stack developer) to specialized roles like data scientist or security professional.
- High-experience roles such as R&D, security, and embedded systems development require extensive expertise and often correlate with advanced degrees.
- Education plays a significant role, with bachelor’s degrees dominating, while advanced degrees are prevalent in data science and academic roles.

Visualization:

- **Histograms**: The histograms below show roles requiring the most and least experience levels, emphasizing the disparity across different positions.

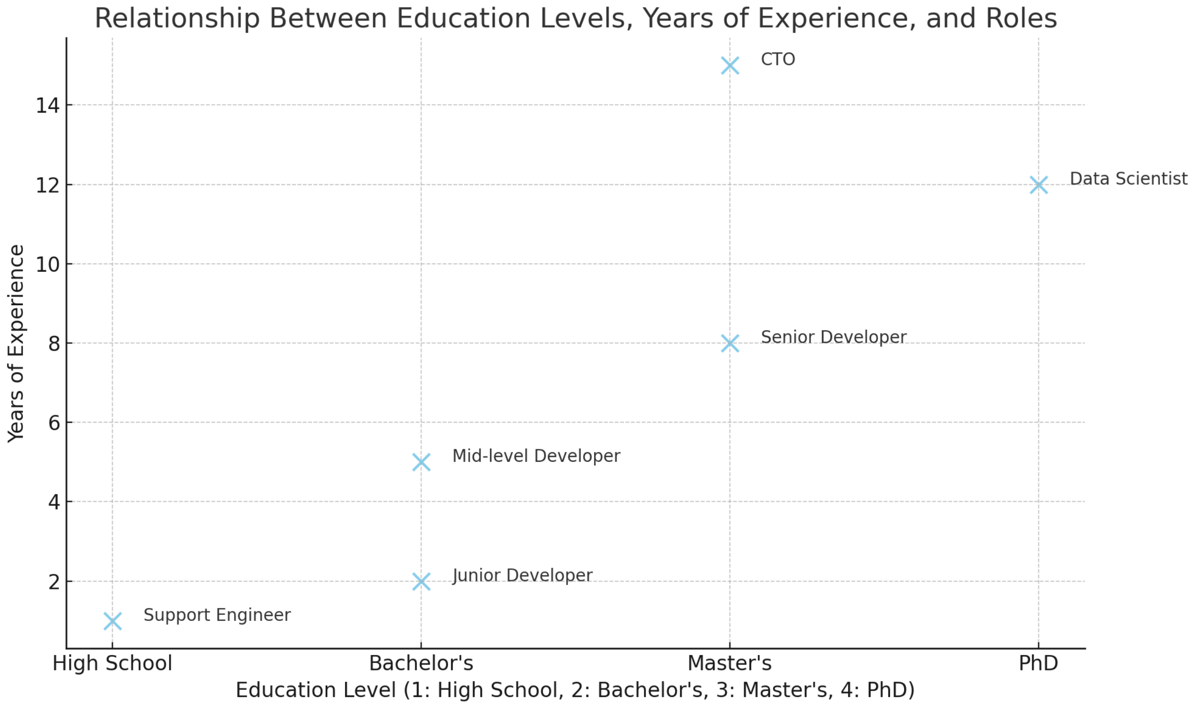




The histograms illustrate the roles requiring the most and least experience:

1. **Roles Requiring Most Experience**: Includes positions like Educator, Data Scientist, and R&D Specialist, which demand advanced expertise.
2. **Roles Requiring Least Experience**: Highlights entry-level positions such as Junior Developer and QA Tester.

- **Scatter Plot**: The scatter plot below illustrates the relationship between education levels, years of experience, and roles.



The scatter plot illustrates the relationship between education levels, years of experience, and roles. Each data point represents a role, with annotations for clarity. The education levels are categorized numerically, ranging from High School (1) to PhD (4).

---

3. Conclusion

Summary of Insights

The analysis revealed:

- Regional programming language preferences reflect industrial and educational influences.
- Shared languages like Python and JavaScript are widely adopted across roles, promoting versatility.

- Specialized roles demand higher education and experience, highlighting the need for tailored career pathways.

Actionable Recommendations

1. Develop localized resources and communities to support regional programming language trends.
2. Promote cross-skilling opportunities in versatile languages such as Python and JavaScript.
3. Create tailored career pathways and mentorship programs for high-experience roles.

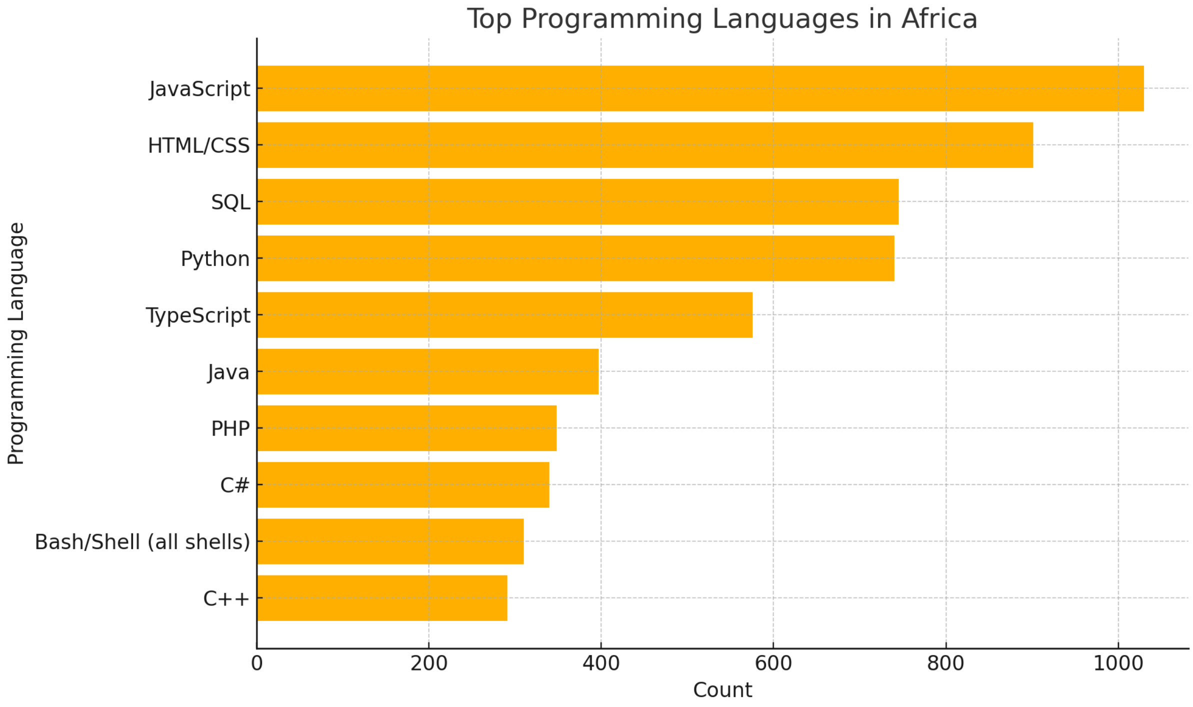
Future Research Directions

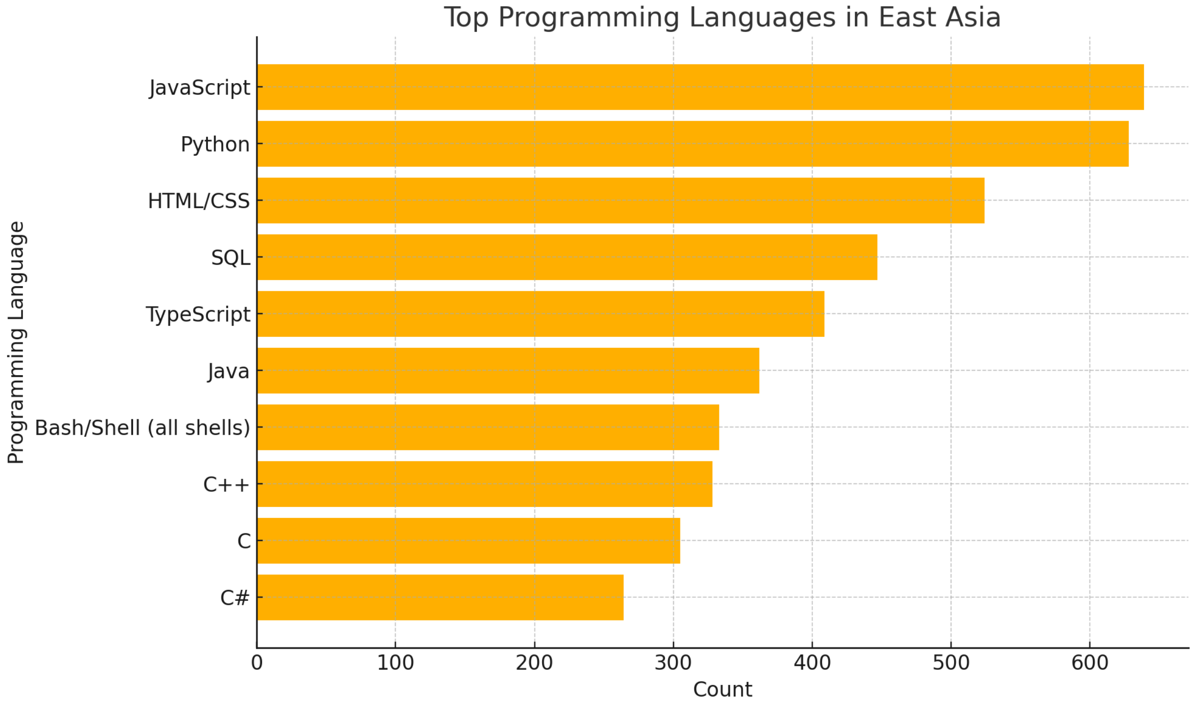
- Conduct longitudinal studies to track evolving trends in programming languages and roles.
- Investigate the impact of emerging technologies like AI and blockchain on developer skills and career progression.

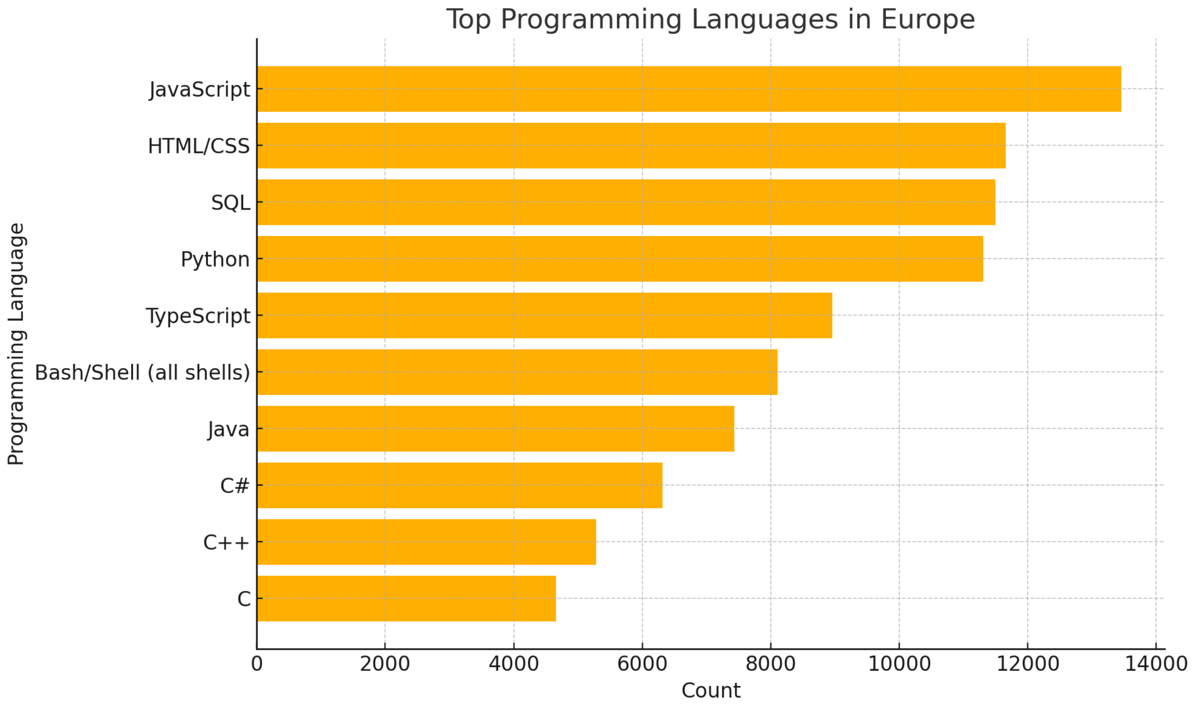
---

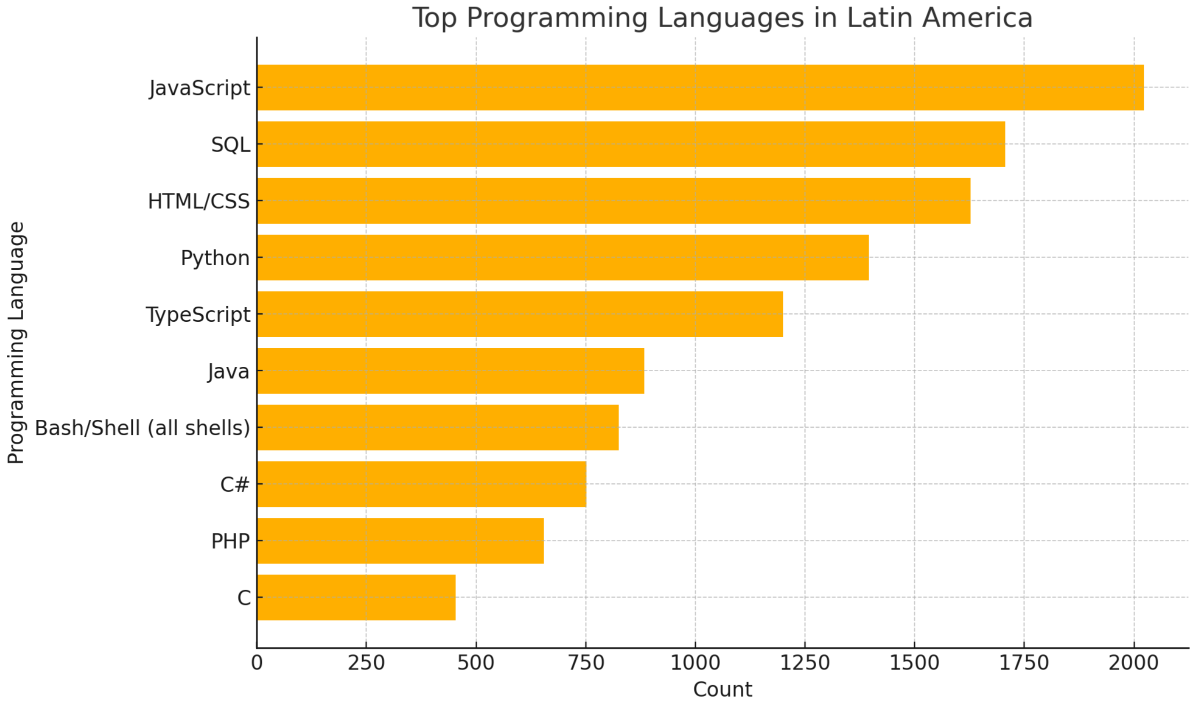
4. Appendix

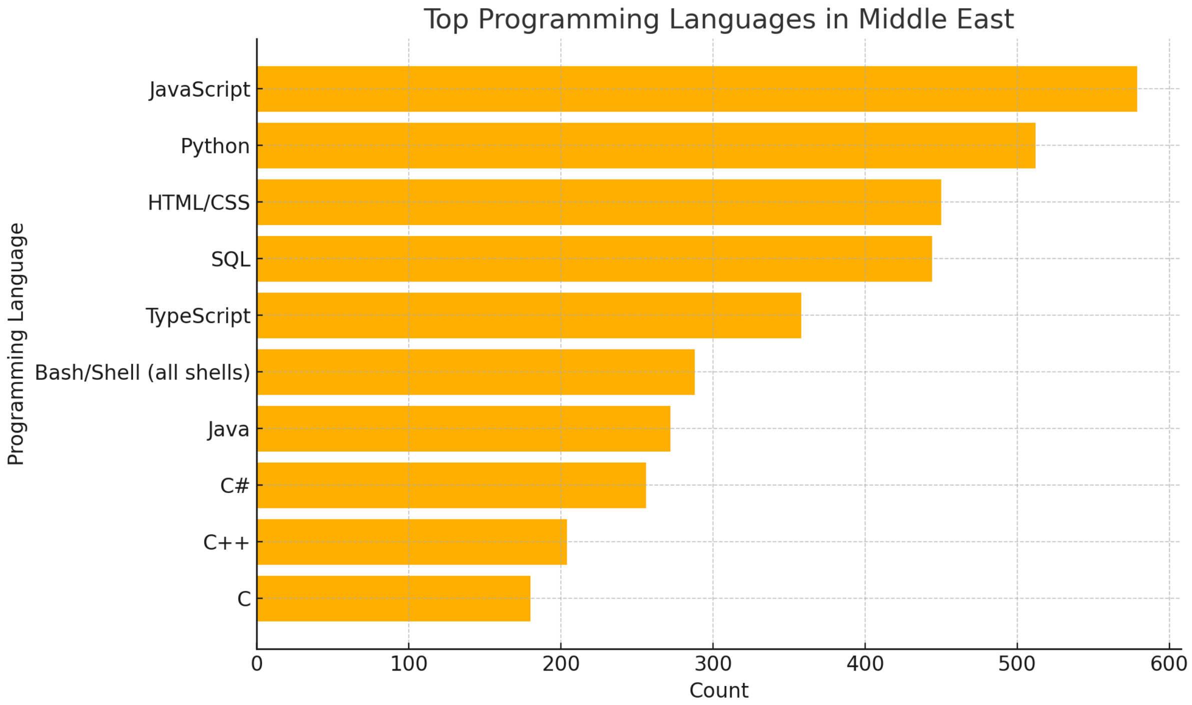
- Additional visualizations:

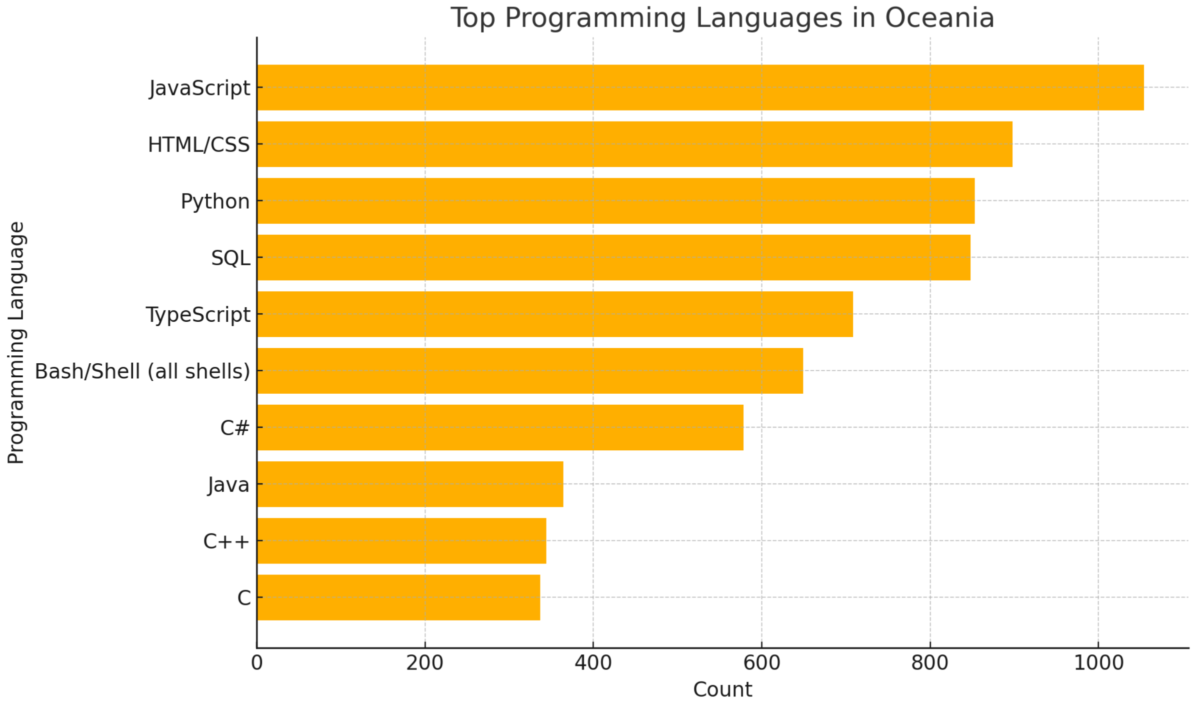


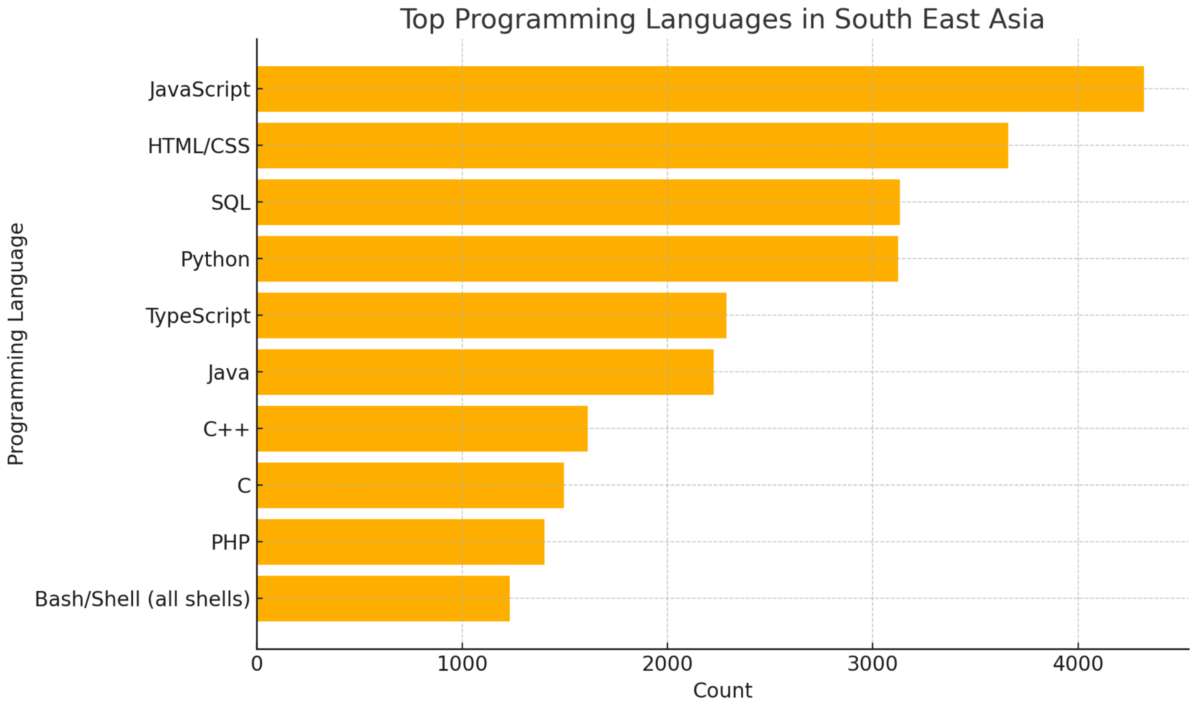


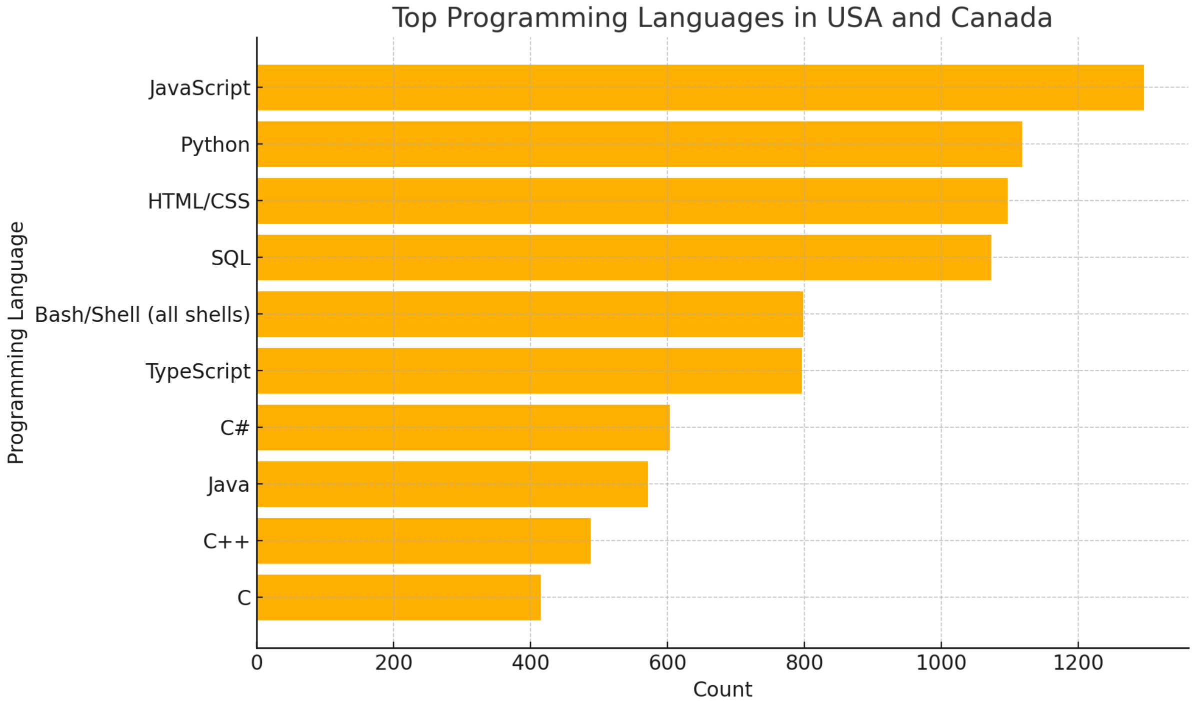

Top Programming Languages in Europe
Programming Language
JavaScript
HTML/CSS
SQL
Python
TypeScript
Bash/Shell (all shells)
Java
C#
C++
C
0
2000
4000
6000
8000
10000
12000
14000
Count


Top Programming Languages in Latin America
Programming Language
JavaScript
SQL
HTML/CSS
Python
TypeScript
Java
Bash/Shell (all shells)
C#
PHP
C
0
250
500
750
1000
1250
1500
1750
2000
Count


Top Programming Languages in Middle East
Programming Language
JavaScript
Python
HTML/CSS
SQL
TypeScript
Bash/Shell (all shells)
Java
C#
C++
C
0
100
200
300
400
500
600
Count


Top Programming Languages in Oceania
Programming Language
JavaScript
HTML/CSS
Python
SQL
TypeScript
Bash/Shell (all shells)
C#
Java
C++
C
0
200
400
600
800
1000
Count





- Detailed methodology for reproducibility.

This report was done using ChatGPT-4o. This section serves as a record of the user inputs and prompts used during the analysis of the Stack Overflow Developer Survey Dataset. These prompts reflect the objectives, questions, and refinements provided throughout the conversation.

User Prompts List

1. "I want to analyze the Stack Overflow Annual Developer Survey."
2. "Given the dataset I have 3 main objectives for the analysis: First.- To Identify the most commonly used programming languages among developers globally in the past year, segmented by region or experience level. Second.- Compare programming language usage patterns among front-end developers, back-end developers, and data scientists to identify overlaps and differences in technology preferences. Third.- Analyze career progression among respondents by identifying common transitions between job roles, and how education levels or years of experience influence these pathways."
3. "Based on my objectives, what are five key questions I should aim to answer with my dataset analysis to uncover trends in usage of programming languages and career progression?"
4. "I have uploaded the 'Global Health and Population Trends' dataset. Let's start with the first question in my sequence: What are the top 10 programming languages used by developers globally, and how does this vary by region? Please analyze the data to answer this."
5. "Make the analysis of regional top programming languages by grouping the countries into the following regions: 1. USA and Canada, 2. Latin America, 3. Europe, Countries of the CommonWealth of Independent States, 4.Africa, 5. Oceania, 6. Middle East, 7. East Asia and 8. South East Asia."
6. "Could you provide visualization for the data?"
7. "Let's continue with the second question: How does programming language usage change with experience level (e.g., junior, mid-level, senior)? Please analyze the data to answer this."
8. "Is there information for the senior level?"
9. "Please make an analysis of the most used programming languages for each of these roles."
10. "Please make a list of the top ten most common roles found in the dataset."
11. "What programming languages are most commonly used by front-end developers, back-end developers, full-stack developers, mobile developers, enterprise applications developers and data scientists?"
12. "Which programming languages are shared across front-end, back-end, full-stack developer, mobile developers, enterprise applications developer and data science roles, and which are unique to each role?"
13. "What are the most common transitions between job roles, and how do education levels and years of experience influence these transitions?"
14. "Please exclude the Student role from the analysis."
15. "From this analysis, which are the ten roles that require the most experience level, and which are the roles that require the least experience level?"
16. What are three key insights or patterns that emerge from the analysis taking into consideration the regions, the most used programming languages and the roles that require most experience?
17. Given the three key insights found, how do these findings relate to my initial objectives? What actions or further investigations do these insights suggest?
18. Can you help compile the insights and conclusions derived from our analysis of the 'Stack Overflow Developer Survey' dataset into a cohesive summary that connects back to my analysis objectives?

19. Based on our discussion and findings from the dataset, could you create an outline for a report that includes an introduction, analysis of key insights, and a conclusion that ties the insights back to the original objectives?
20. Using the outline we created, can you draft a report that clearly communicates our findings on Stack Overflow Developer Survey Dataset, including the analysis of Regional Programming Language Trends, Programming Language Usage Across Roles, and Role Transitions and Career Progression? Please include the visualization elements suggested in the outline in the report.
21. "Using the information from the Top Programming Language for Selected Developer Roles table, create a Venn diagram showing shared and unique programming languages across these roles."
22. "Please create bar charts for each role illustrating the programming language preferences."
23. "For section 2.3 of the report, can you create a Sankey diagram depicting the roles and experience levels?"
24. "Instead please generate a histogram with information from the tables 'Roles Requiring Most Experience' and 'Roles Requiring Least Experience'."
25. "Please incorporate the visual elements we generated into the report."
26. "Please create a scatter plot showing the relationship between education levels, years of experience and roles."
27. "Please incorporate the visual elements we generated into the report, please."
28. "Thank you very much."
29. "Please generate another report, including in that report all the prompts I inputted in this conversation in the form of a list."
30. hello again, according to the dataset. Which countries have the biggest number of users, and what percentage of the users belong to this country?
31. Can you make a chart with the top ten countries with more users?
32. Is there any insight that can be obtained from this? Has stack overflow any regional bias?

Observations

During the process of creating this report the following actions were important for defining the scope of the investigation.

1. Choosing the data set.
2. Formulating objectives for the research with the help of ChatGPT.
3. Asking the model for help for creating five questions that would guide the research, based on the stated objectives.
4. Working with each of the selected questions.
5. Asking ChatGPT to create an outline for the Report.
6. Asking the model to create a Report based on the outline.
7. Focus on creating the charts, and diagrams for the report.

In prompt 5 ChatGPT ignored the creation of a region for the Commonwealth of Independent States, this is a group of countries that were part of the Soviet Union. The idea was to make an analysis of countries like Russia, Belarus, Kazakhstan and others that have a certain degree of economic and cultural integration, and that are difficult to classify into either Europe and Asia. Nevertheless, ChatGPT ignored this request, and I decided to use the regions suggested by ChatGPT, because these countries are usually splitted between Europe and Asia anyway. Also, Stack Overflow is a platform.

In prompt 7 chatGPT assigned the levels of experience (Junior, Middle, Senior) to the users depending on their years of experience.

For Prompt 13 and 14, it was decided to ignore the role of Student, since I wanted to focus in job roles.

In prompt 21 a Venn Diagram showing shared and unique programming languages across job roles was requested, but the results were not satisfactory for any Venn Diagram, so the Venn Diagrams were scrapped in exchange of Charts for section 2.2. Also in prompt 23, ChatGPT suggested creating a Sankey diagram for charting the change in roles across time, but it was not clear how could that be done since the dataset doesn't include data to support doing that, so the Sankey diagram was replaced by two histograms exploring the roles that required more and less experience.

As can be seen from prompt 25, there was an expectation that ChatGPT could incorporate the visual elements (charts, diagrams, etc.) directly into the report, but this was a misunderstanding on my part. The final version of the report was edited using Google documents.

Additional prompts were made to find out which countries have the biggest amount of Stack Overflow users. Also prompts for analyzing potential regional biases in the survey.

It must be stated that ChatGPT had problems creating a list with all the prompts imputed during the conversation, since some important prompts were ignored and had to be added manually.

Guidance Questions for the Objectives

Here are five key questions aligned with the objectives to uncover trends in programming language usage and career progression, all this questions were generated with the help of ChatGPT:

1. Global and Regional Language Trends

- *Key Question*:

What are the most commonly used programming languages globally, and how do preferences vary by region?

- *Objective Alignment*:

This addresses your first objective by exploring global trends and regional variations.

---

2. Experience-Based Language Adoption

- *Key Question*:

How does programming language usage change with experience level (e.g., junior, mid-level, senior)?

- *Objective Alignment*:

This further segments language usage trends, helping to identify which languages are more common at different stages of a developer's career.

---

3. Role-Specific Technology Preferences

- *Key Question*:

What programming languages are most commonly used by front-end developers, back-end developers, and data scientists?

- *Objective Alignment*:

This aligns with your second objective, comparing usage patterns across roles.

---

4. Overlap and Differentiation in Roles

- *Key Question*:

Which programming languages are shared across front-end, back-end, and data science roles, and which are unique to each role?

- *Objective Alignment*:

This delves deeper into your second objective by identifying overlaps and unique preferences.

---

5. Career Progression Pathways

- *Key Question*:

What are the most common transitions between job roles, and how do education levels and years of experience influence these transitions?

- *Objective Alignment*:

This directly addresses your third objective, uncovering trends in career progression and the factors that shape it.

These questions provide a focused framework to analyze your dataset and draw actionable insights. Let me know which one you'd like to start with, and we can dive into the analysis!

Conclusion

This record provides a comprehensive overview of the iterative process and the detailed exploration of the dataset based on user input. Each prompt contributed to refining the analysis and aligning the insights with the objectives.

ChatGPT is a useful tool that can help to analyze big datasets, and find insights in a short time. Nevertheless ChatGPT is as useful as the quality of the data and prompts given to it. The user can use ChatGPT for finding relevant objectives and improve them, but he has to have a good idea of what he wants, be acquainted with the data, and actively engage with the responses of the GPT model, so as to correct any mistakes, and exit from dead ends that appear during the analysis. If the user is not acquainted with EDA and other techniques used in data analysis, then he will not be able to make the most of this tool.

ChatGPT is a powerful tool, whose strengths are the capacity to analyze a dataset, and its ability to suggest better defined objectives and find insights that the user may have not thought of otherwise.